Quantitative Analyses in Global and Area Studies using Graph-based Filtering of Heterogeneous Catalogue Data

Thomas Efer, Ninja Steinbach-Hüther

Computer Science Department | Graduate School "Global and Area Studies"
University of Leipzig
04109 Leipzig
efer@informatik.uni-leipzig.de
ninja.steinbach-huether@uni-leipzig.de

Abstract: In this contribution we describe the objective and technical considerations of an ongoing early-stage interdisciplinary collaboration at the University of Leipzig for the preliminary part of a quantitative analysis within a project of the Graduate School "Global and Area Studies".[1] This project is located at the meeting point between approaches stemming from the e-Humanities and approaches coming from the Global and Area Studies. Our intention is to provide first insights into new, data-driven methodology for bibliography-based analyses. We want to show the utility of recent technology for data modeling and management in the Global and Area Studies.

1 Research Questions

Africa seems to have become a focus of higher academic interest in various disciplines among which the Area Studies play a considerable role in this debate. Hence, if interest in a certain region grows significantly, we are most often confronted with the need for publications from and on this region. But questions on the presence and reception of African (academic) books and their place in academic debates are often answered by emphasizing their marginal role and their quasi non-existence on a global level [Bg06]. Nevertheless, despite critical statements about the global ignorance towards African academic literature, we seem to lack an all-encompassing overview over its quantitative status and how this has developed for the last five or six decades in certain regions and fields of research. Against this background, this contribution is placed in the broader scale of a project of the Graduate School "Global and Area Studies" that exemplarily analyses the presence of African academic literature in Germany and France in historical perspective.[2] The key question behind this project is whether we are dealing with a

[1] The methodological approach presented here is part of a larger PHD-project to be submitted by N. Steinbach-Hüther at the Faculty of Social Sciences and Philosophy, University of Leipzig, for which the methodology will be developed further. The results presented here are a first step for the larger study.

[2] This rather European angle and perspective can be explained by the fact that direct publishing in European or US American publishing houses as part of the global North has been and still is very common despite

growing attention to Africa and the knowledge that is produced by its academics or whether African knowledge is still situated at the margins of global knowledge production.[3]

The project theoretically follows the concept of cultural transfer as was established by Espagne, Lüsebrink and Werner in the French 'German studies' as part of the cultural studies and an alternative to the traditional comparatistic [Es06]; [KS03]; [Lü08]; [Mi01]; [WZ03]. The methodological work combines various methods to accomplish an all-embracing reconstruction of the whole transfer process in both a quantitative and a qualitative sense [Es12].

The scientific aim of this project is firstly to examine the quantitative presence of African academic literature from the Social Sciences with the Humanities also included, but the Life Sciences are excluded. In this work we concentrate on the French part of the analysis only.[4] Therefore, we will present the preliminary methodological approach to a rather complex dataset that was compiled by the *Bibliothèque nationale de France (BnF)* and that is based on its bibliographical catalogue[5]. By evaluating the entries we will accomplish the aim of getting bibliography-based quantitative results as the foundation for further quantitative and qualitative research in the Global and Area Studies.

Building on that primarily work, it can be analysed in a second step to what extent "academic Africa" has finally paved its way to the French book market and correspondingly may have become an issue of academic debates, discussions and discourses. While the first question is rather quantitatively oriented, the second is based on its results and approaches the results in a qualitative way. In a nutshell: based on the analysis, it will be dealt with the question of inclusion versus exclusion of African knowledge from the academic knowledge debate. That is why this work follows two different aims that can be accomplished by different methodological strategies only.

What is especially interesting regarding this methodological handling of the dataset is the fact that the research question could not be addressed and fulfilled the way it is if it had been for leaving the analyses of such a huge dataset to the conventional way of doing computations and the very restricted set of traditional digital tools that can be found in usual office software. The insight should be carried out beyond the reach of Global and Area Studies that custom-tailored digital tools, forged in an interdisciplinary setting, can enable broader research questions. The e-Humanities pose a lot of interesting questions especially to the knowledge-centered parts of the Computer Sciences. In our case the question prevails how to model library data in a most fitting way for specialists of certain fields other than the library sciences. How to pave the way and build a common technological ground for similar projects?

---

numerous publishing initiatives during the last two decades to publish literature in African countries themselves.

[3] The research question is explained in detail at http://www.uni-leipzig.de/~ral/gchuman/en/units/graduate-school-global-and-area-studies/transnationalization-and-regionalization/doktoranden?tx_wecstaffdirectory_pi1[curstaff]=86

[4] For some of the results of the German part of the analysis see [MSH14b].

[5] http://catalogue.bnf.fr/

2 Datasets and Challenges

The raw datasets were compiled by the BnF from their catalog repository and according to four rather broad criteria:

- at least one of the authors of a publication has to be somehow connected to the African cultural region according to the denoted country or language code in the general catalogue ("catalogue général")
- the publications linked with these people became relevant only if published in the form of books/monographs
- the publications had to be published after 1950
- the publications had to be published in France[6]

To accomplish a corresponding filtering, the BnF first compiled a list of authors who have been assigned an African nationality and one of those for whom an African language skill cannot be ruled out[7]. For those authors of potential interest all those books (including their bibliographical data) were selected in a second step, for which the previously extracted authors acted as primary or (one of the) secondary authors. Both lists (51.266 authors and 64.574 books) were made available through a data set which is based on the *INTERMARC format*[8]. This textual format works with a "zone"-based mechanism for structuring entries[9] which can hold by themselves multiple values which are indicated by different "splitter sequences" like "$a". In the Excel-based output format some of those multi-values were already resolved into their own columns, while some remained combined. The individual records originate from different internal systems. This leads to the heterogeneous nature of the dataset with different levels of data granularity and completeness.

The retrieved lists use numerical identifiers for publications and authors but do not provide identifiers or other special notions for other entity types such as cities or publishers. Those are compiled on a textual level only and seem to have been manually collected without being fully normalized since spelling and expression variants persist throughout many fields.[10]

From this originates the need to account for the various common tasks of heterogeneous data integration such as data cleaning and normalisation.

---

[6] Excluded are books which are not published in France but that were bought by the BnF as well as books published in francophone countries other than France that also fall under the National Library's collective order (e. g. some African countries, (francophile) Canada and Switzerland. Some additional African authors (those with French citizenship that otherwise would have been excluded from the analysis) could be found via their language code.

[7] That includes people with an African language code (likely Africans) as well as people generically classified as "multilingual".

[8] http://www.bnf.fr/en/professionals/intermarc_format_eng.html

[9] Zones are fixed 3-digit numbers that denote a common field type. An example for possible sub-values for such fields and their delimiters can be seen here for zone 260, the publisher's address: http://www.bnf.fr/documents/pb-RIMB08_260.pdf

[10] This leads to the question about who actually collects/makes and gives the codes in national bibliographies and in libraries respectively.

The BnF provides their whole catalog dataset as linked open data in the RDF format[11] and explains the process of *Linked Data* publication in depth in [SWMD13]. This complete data dump (containing over 38 GB worth of uncompressed XML files) was loaded and inspected. It showed, that this dataset does not constitute a real advantageous alternative over our Excel files for several reasons:

- Dealing with the inconvenient size of the dataset in line with the need for a pre-filtering according to the above-mentioned criteria would have been an additional task for us.
- We saw problems of communicating the "raw" data model which consists of vocabularies from 20 different namespaces to non-computer-scientists in order to define a common starting ground. It proved too complicated to find clear transformation steps towards the research questions in this setting.
- While multi-valued fields are properly split (which gives us an evaluation basis for our splitting), no substantial normalisation steps seem to be taken. Full text field content stays full-text. Orthographic variation is not resolved.
- Semantic enrichment and linking in the RDF data set do not focus on the domain of our research question.

The last point is really important: While there exists for example a comprehensive geographical linking of books and places on the FRBR[12]-level of "Manifestations", its semantics seem to lie solely in the description and referencing of the work's subject matter and not in the organisational aspects (e.g. the publication location of the edition) that we are interested in. Even more interesting entities such as publishers are not linked or assigned a unique identifier.

But even with an extended linking of all relevant aspects, there could remain uncertainties concerning the achieved accuracy and completeness of the "semantification" process that may render it impractical to build accurate scientific methods on top of such a (semi)automatic external preprocessing. After all it may be better to perform such crucial steps by oneself to keep full control. So we refrained from using the RDF data set in our analyses. Yet it can be seen as a remarkable and much appreciated step of the BnF, that this substantial amount of internal data is provided within such an open technological and legal framework.

3 Implementation

For the cleaning and normalisation of entries as well as their reliable interlinking according to common properties there exists a plentitude of possible technological resources but still no definitive methodology or toolset. The consumer-specific way of handling and enriching heterogeneous data is in our case quite comparable to the idea of "self-service Linked Government Data" as presented in [MCP12], where raw publicly

[11] http://data.bnf.fr/semanticweb-en, subject to a special license: http://data.bnf.fr/licence
[12] Functional Requirements for Bibliographic Records, see http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

available data sources are enriched, linked and re-published by the data consumer. Its proposed workflow is based on Open Refine[13] (formerly Google Refine), a very capable tool for data cleaning and linking.

Such standard methods and tools are a big help to fix the issues in raw datasets. But the interconnected nature of a bibliographical dataset[14] renders them less useful, as they see the datasets primarily as independent tabular structures. Yet the biggest problem with such tools is, that all enhancement operations like normalization, splitting and merging as well as re-linking are irreversible and permanent (with the exception of linear undo/redo functionality), and lack options to properly document the intention behind them. This poses their most prominent disadvantage for an explorative academic scenario where hypotheses have to be tested thoroughly and in many cases have to be abandoned later on in favour of new ones, without starting over again on all the other aspects, too.

Besides that category of programs, specialized transformation and integration tools for library (meta) data exist, see for example Metafacture[15]. Their focus lies in the definition of static transformation workflows for changing datasets, whereas we need a flexible explorative system for a static dataset where we mainly aim at filtering data than to completely transform all entries into another format.

So we decided to build a completely independent toolbox to be able to design all aspects according to the scientific quality requirements for the Global Studies. The toolbox consists of a data store and integrated process chain and will be extended with a web-based visual analytics interface. While developing, a prototype-driven process allows for an exploration of the peculiarities of the dataset prior to fixing the complete strategy for approaching the research questions.

The process consists of several steps. Pre-Processing consists of converting all record entries to single linkable representants yielding the original field values. Then the field values where fixed parsing can be applied (like multi-valued fields) are transformed to represent specific human-interpretable values of determinable semantic. Finally there is an iterative process of creating machine-readable, unified, normalized and queryable values. The needed data store has to support short feedback cycles (preferably without demanding a fixed schema), it should be capable of modelling a “natural” networked view on the data so that normalization is not only lexically motivated (e.g. truncated full text search) but can derive new identity of entities and semantic concepts (publishers, genres and the like) by adding new entities to the datastore.

According to good experiences in previous smaller projects we chose the Property Graph model as fitting representation and the de-facto leading Graph Database *Neo4J*[16] as our backend. We decided to use *JRuby*[17] as programming environment since it enables us to

[13] http://openrefine.org/
[14] Different entity types such as people, places, subjects and organisations all contribute to the dataset in an individual way.
[15] https://github.com/culturegraph/metafacture-core/
[16] http://www.neo4j.org/
[17] http://jruby.org/, a Ruby runtime based on the Java Virtual Machine

use Java libraries (such as an embedded Neo4J engine) while we can at the same time benefit from Ruby's concise syntax and other handy options, e.g. the ability to extend the language with dynamic programming concepts to act as *Domain Specific Language*. The datasets are in the *Office Open XML* Format and can be read with the library ("ruby gem") *roo*[18]. Since this mode of accessing the data is cell-based and therefore very slow we decided to first transform the Excel files into a raw tabular JSON format to have a fast way to easily load the table into memory on demand.

The conversion of those raw tables into a graph then consists of multiple steps. The splitting of multi-fields and linking of books and authors via their IDs is a purely rule-based mechanism. Then follows an "*interdisciplinary explorative rule mining*" process, where a growing and alterable list of steps is employed to record transformations . The effects of those transformations can then be assessed in an analytics frontend. The work is still ongoing and can therefore only be briefly outlined at the moment:

The existing graph gives insights into the distribution of several field values, link patterns and other properties. The researchers have to decide if such patterns represent

- useful filter criteria,
- candidates for the merging of datasets,
- or just insignificant statistics

Based on that decision a set of graph queries for reading and transforming data and then altering the graph is built and successively extended. When some rules are later on deemed unsuitable, a reprocessing has to take place. We managed to improve the performance of Neo4J for that task significantly from hours to minutes by using "unforced" transactions[19]. This speed makes the re-processing of changed rulesets quite convenient. In addition, a partial re-processing from fixed-state copies can be introduced.

Using these query collections we can not only alter the existing entries but also add new entries for discovered entities and link them to the basic data sets according to field values. It will become far easier to then find a place of publication, a publishing house, a translation or a book of a particular genre if all variables in the writing have been attributed a new single qualifier which at the same time combines all the former qualifiers to one (such as one graph node for the differently written names of one publishing house). This process goes hand in hand with data exploration and gradually improves the dataset while keeping track of all performed actions. The goal behind this setup is to have a flexible environment not only to support our own research purposes but also to be able to take part in larger efforts of *Knowledge Federation*. So we took care to ensure that our graph based model can be easily exported into popular formats such as RDF or Topic Maps. For best practices in that field see for example the also JRuby-based approach in [BJSM10].

[18] https://github.com/Empact/roo
[19] http://components.neo4j.org/neo4j-kernel/1.9.M01/apidocs/org/neo4j/kernel/TransactionBuilder.html#unforced()

4 Towards a new Methodology

When we consider a simple example where we want to determine if a book is a novel or not, we can imagine plenty of different indicators that contribute to that decision: There are *directly accessible properties* such as the word "roman" attached to the title field. This is a quite common practice but still gives results that are far from being complete. Additionally we can include *indirect approximations*. If an author has written ten books and nine of them are already classified as novels we may want to interpolate this class for the tenth book too. This may apply even more for publishers specialised in a certain field. Finally there is the option to define certain *context-sensitive rules*: One could hypothesize that books with more than 2 authors are most likely not novels. This hypothesis can be tested beforehand using the already classified works and then be applied to define exceptions to other rules, revert a previous classification, notify the user or the like.

In general such a combination of several incomplete heuristics is a valuable tool in an explorative filtering workflow. The *stepwise refinement* of classifications gradually reduces the uncertainty of the final quantitative aggregations while it possibly also narrows down the selection of objects of interest to a manageable portion for manual qualitative analyses. To support this (methodologically rather experimental) course of action, there must be flexible tools and data-centric prototypes that are developed in close dialogue between Humanities scholars and computer scientists.[20]

All taken steps (data cleaning and normalisation followed by classification, inference and filtering and finally querying and aggregation) should be available through an integrated toolbox. In most cases all of them can be expressed in terms of simple graph queries and value assignments. Those queries *make explicit* and therefore *document* the process of data transformation and selection. They are ensuring reproducible final results while providing flexible means for further revision and alteration of intermediate processing steps – and that at purely computational cost and not with manual effort. By introducing such tools and data processing methods into the Humanities and Social Sciences, new questions can be posed and successively answered through data analysis. As already stated above, this methodological toolset is beneficial for quantitative studies as it makes possible comprehensive inquiries on large datasets as well as for qualitative studies for which proper examples can be picked in a representative way.

5 Preliminary results

The conducted work up to this point is characterized by an explorative approach to get to know the datasets better and to decide on the applicability and fitness of quantitative methods in the context of the research questions. Can we identify information on sticking points, continuities, discontinuities and breaks in the period of the study,

[20] If sufficiently funded and after evaluation, this may then lead to being able to reproducing the method in other interdisciplinarily oriented projects. Similar research questions could henceforth be addressed with an easier methodological starting point.

especially with regard to the chosen thematic, geographic and temporal profiles of the published material but also with regard to personal, lingual and editorial dominances?

For example how big is the influence of certain publishers or institutions in the whole publishing process? What kind of publishing houses supported the publishing of African academic literature throughout the years and which role did publications in their original language have in this process in comparison to translations? It will also be interesting to get to know which books from what kind of authors were published and where the authors came from. Can we identify changes in thematic, lingual and geographic orientation during the time period for example? In this respect, results of the quantitative analysis will show the conjunctures of the whole transfer process in matters of the progress in time, the chosen topics and the privileged original languages but also relating to the authors themselves and further actors such as translators and publishers.

This is directly linked to new methodological questions: Should the analyses be split by time periods, for example, to show geographical, lingual and thematic dominions as well as correlations between them? Regarding figure 1 that shows how *L'Harmattan* has constantly increased its share in the total numbers of the books in the given list in comparison to *Karthala* that seems to have a rather constant share of the books, another methodological question follows: Should large editorial players, that quantitatively determine the numbers, be analysed individually later on with respect to the research question to prevent the insight into representative results from being altered by internal publisher policies and positions towards African authors? It is an important lesson that such analyses can only grow to full potential when guided and directed by explorative data analyses. If conducted in an interdisciplinary way this will become even more fructuous.

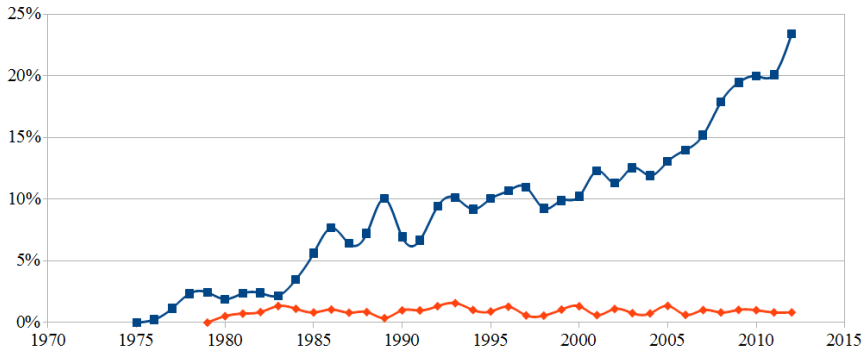

Figure 1: Percentage of all included books which are published by L'Harmattan (blue) and Karthala (orange), calculated by merging over 60 different renderings and typos of the publishers' and their spin-offs' names

In this project phase graph queries have proven to allow for quick interaction with the enhanced dataset, giving smaller result sets in mere milliseconds while returning graph-global aggregations after few seconds. The included full text indices for node properties allow an almost instantaneous display of time series for the occurrence of certain keywords in titles, notes and other such textual fields. This runtime performance allows

for the later creation of query interfaces to power a user-friendly interactive toolbench in the spirit of Visual Analytics, see [KKEM10].

About 45% of all listed books have a DDL class assigned. An analysis of the actually used categories (see Figure 2) shows as expected an uneven density and depth throughout the categories.[21]

By first intuition the categorisation space can be quite easily divided into segments belonging to one of the following three classes:

- definitive rejections (e.g. works from the natural sciences or works of fiction)
- definitive acceptance of categories (because they obviously lie in line with the research questions)
- a few borderline cases that soften the "accept"-category (e. g. books classified within the spectrum of medicine but oriented to the social aspects of the field rather than to the medical) or the "reject"-category respectively (e. g. books classified as belonging to the arts but treating its history and interpretation rather than being illustrated books)

All should be easily recognisable and determinable and form larger regions of common classes. Yet in practice almost every category yields strong potential to belonging to the "borderline" class, rendering the DDC (although partially very fine-grained) a tool which is still too coarse and detached from the research question as to be a solitary filter criterion. One reason for this is that in the Humanities and Social Sciences there is much impetus to reflect on or to analyze phenomena within a large spectrum of socially relevant topics and particular communities. Therefore they also contribute written (meta)literature to that field which in many cases has to be classified with the same category as its subject matter literature.

What is important to note so far is that in the field of the Social Sciences as well as the Global and Area Studies, it is very advantageous to make use of comparable programs that facilitate quantitative analyses the way it was shown. After all, the results of this analysis will show the most exact number - that is to say the quantity of books that was systematically possible to capture - of the presence of African academic literature from the Social Sciences and the Humanities published in France since the 1950s.

Of course, the results need to be *counterproofed* in a second step according to their relevance for the rather qualitatively oriented second research question but it would not have been possible to get the results in the first place if it had not been for the use of such an informatic tool to handle the sheer amount of data. It goes without saying that it was important to have generated rather too many results in the first place than too little and to therefore check the data in a second step according to its relevance.

[21] Since the DDC combines enumerative and faceted principles some of the emergent patterns may be representations of common semantics and some only co-incident.

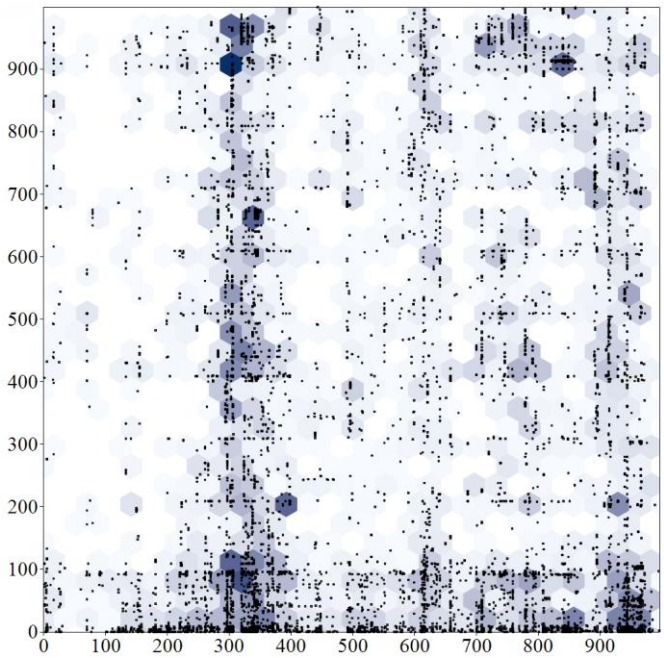

Figure 2: Graphical representations of occurring DDC numbers (first three digits as combined number on x, second three on y; background: hexbin choroplet map for easier density estimation)

It is also worth it to note that the method as such, even if that much effective, reproduces blurs as the search runs go in line with predefined parameters. As was stated above we should ask who actually collects/makes and gives the codes in national bibliographies and in libraries respectively. These codes are assigned manually and therefore according to a judgment of the person who actually completes the data record, which is why we have to keep in mind that this is one of the inevitable but reproduced blurs of the study.

For the computer sciences and their growing engagement in the e-Humanities it is important to take away from such projects the lesson on how crucial it is to offer flexible, normalized, disambiguated and entity-centric models of domain knowledge to researchers alongside semi-automatic methods for their research-centric unification and enrichment.

References

[Bg06] Bgoya, W.: Introduction – Scholarly Publishing: An Overview. In (Mlambo, A., Ed.): African Scholarly Publishing Essays, Oxford 2006; P: 5 ff.

[BJSM10] Bleier, A.; Jähnichen, P.; Schulze, U.; Maicher, L.: The Praxis of Social Knowledge Federation. In (Karabeg, D.; Park J. Eds.): Self-Organizing Collective Mind, Second International Workshop on Knowledge Federation, Dubrovnik 2010

[Es06] Espagne, M.: Jenseits der Komparatistik. Zur Methode der Erforschung von Kulturtransfers. In (Mölk, Ulrich, Ed.): Europäische Kulturzeitschriften um 1900 als

Medien transnationaler und transdisziplinärer Wahrnehmung. Bericht über das zweite Kolloquium der Kommission „Europäische Jahrhundertwende. Literatur, Künste, Wissenschaften um 1900 in Grenzüberschreitender Wahrnehmung“, Göttingen 2006.

[Es12] Espagne, M.: Comparison and Transfer: A Question of Method. In (Middell, M.; Roura, L., Eds.): Transnational Challenges to National History Writing, Basingstoke 2012.

[KKM10] Keim, D.; Kohlhammer, J.; Ellis, G.; Mansmann, F. (Eds.): Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics, 2010

[KS03] Kaelble, H.; Schriewer J. (Eds.): Vergleich und Transfer. Komparatistik in den Sozial-, Geschichts- und Kulturwissenschaften, Frankfurt am Main/New York 2003.

[Lü08] Lüsebrink, H-J.: Interkulturelle Kommunikation. Interaktion, Fremdwahrnehmung, Kulturtransfer, Stuttgart/Weimar, 2nd edition, 2008.

[MCP12] Maali, F.; Cyganiak, R.; Peristeras, V.: A Publishing Pipeline for Linked Government Data. In (Simperl, E.; Cimiano, P.; Polleres, A.; Corcho, O.; Presutti, V. Eds.) The Semantic Web: Research and Applications, LNCS 7295. Springer Berlin Heidelberg 2012; P: 778-792

[Mi01] Middell, M.: Von der Wechselseitigkeit der Kulturen im Austausch. Das Konzept des Kulturtransfers in unterschiedlichen Forschungsrichtungen. In (Andrea Langer, Ed.): Metropolen und Kulturtransfer im 15./16. Jahrhundert, Prag/Krakau/Danzig/Wien 2001.

[MSH14b] Middell, M.; Steinbach-Hüther, N.: La présence de la littérature académique de l’Afrique en Allemagne. In (Espagne, M.; Lüsebrink, H.-J., Eds.): Transferts Culturels (exact title still to be given, to be published in 2014; Paris: Karthala)

[WZ03] Werner, M.; Zimmermann, B.: Penser l’histoire croisée. Entre empirie et réflexivité, in: Annales, 58/1. (2003); P: 7-36.

[SWMD13] Simon, A.; Wenz, R.; Michel, V.; Di Mascio, A.: Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF. In (Cimiano, P. et al. Eds.) ESWC 2013, LNCS 7882; P: 563–577